

Predicting demand in changing environments: a review on the use of reinforcement learning in forecasting models

José Rolando Neira Villar, Miguel Angel Cano Lengua

Facultad de Ingeniería de Sistemas e Informática, Universidad Tecnológica del Perú, Lima, Perú

Article Info

Article history:

Received Jun 12, 2024

Revised Oct 23, 2024

Accepted Nov 19, 2024

Keywords:

Adaptative algorithms

Artificial intelligence

Concept drift

Demand forecasting

Reinforcement learning

ABSTRACT

This systematic review, carried out under the PRISMA methodology, aims to identify how reinforcement learning has been used in demand forecasting, distinguishing the problems they are trying to overcome, recognizing the algorithms used, detailing the performance metrics used, recognizing the performance achieved by these models and identifying the business sectors in which it has been developed. Studies from all sectors were considered to expand the search range. A total of 24 articles were qualitatively analyzed, and the main results were that reinforcement learning has been used mainly for the selection or dynamic integration of the best predictors from a base of them to adapt to changing environments; whereas forecasting in volatile and complex environments is the main issue addressed; whereas Q-learning (QL), deep q network (DQN), double deep q network (DDQN), and deep deterministic policy gradient (DDPG) are the most widely used algorithms; and that, finally, the sectors of electric power, thermal energy, transport and telecommunications are the sectors where this type of forecast has been developed. Finally, given that all the models studied lack mechanisms for detecting concept drift, a new use of reinforcement learning for this purpose is proposed.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Miguel Angel Cano Lengua

Facultad de Ingeniería de Sistemas e Informática, Universidad Tecnológica del Perú

Jr. Natalio Sanchez 125, Lima, Perú

Email: mcanol@unmsm.edu.pe

1. INTRODUCTION

Previous research [1], [2] has established that accurate demand forecasting is crucial for the efficiency and proper development of fundamental business activities such as the design and management of production processes, the determination of purchasing and inventory levels, the projection of cash flow and capital needs, as well as planning for hiring and training staff. In addition, it has been identified that having advanced capabilities in demand forecasting provides a significant competitive advantage, avoiding serious problems such as excess inventories, shortage of supplies for production, high labor costs, and loss of reputation. Due to their enormous importance, in recent years innovative forecasting methods based on artificial intelligence models have been proposed, which have recently been collected in literature reviews like [1], [3]. However, in these studies models based on reinforcement learning are conspicuous by their absence.

Despite the importance of these innovations, the challenge of forecasting the demand in many markets remains enormous. The nonlinearity, volatility and dynamism of the characteristic fluctuations of demand in some sectors such as electric power and transport [4], [5], the presence and juxtaposition of demand variability at different time scales in sectors such as telecommunications [6], the large number of explanatory variables of demand for various products that can produce very complex patterns such as in the

semiconductor distribution sector [7], they call for forecasting methods that are able to adapt quickly to changes in the market situation. Furthermore, these significant changes in demand patterns have a very important implication for the widespread forecasting models based on supervised learning algorithms. These kinds of algorithms are trained only once with historical data and, once the training is complete, they are used to forecast with data that the model has not seen before. It is assumed that the relationship between the input data and the output variable will always be like that found in the training data. However, this is not necessarily the case: when this relationship changes over time we have what is called concept drift. When this occurs, supervised learning models lose their precision and eventually need to be retrained or updated with new data to avoid becoming obsolete [8]. To address the problem of forecasting in changing contexts, in [9] the authors propose the dynamic assembly of a broad base of different types of predictors that are joined by weighted weights found based on the recent performance of each predictor. In this way, the predictors that are part of the assembly change to give rise to those that best capture the fluctuations that occur in the market. However, despite its benefits, this model lacks the means for timely detection of concept drift to make the necessary updates to its predictors [9].

In line with the above, Gama *et al.* [8] established that a genuinely adaptive model must have a change detection mechanism that can determine the presence of concept drift to act accordingly. Among the possible options for this mechanism, they propose the use of statistical control charts on forecast errors. The underlying idea is that in the face of significant changes in the relationship between the input variables and the predicted variable, predictors trained on data that are no longer in force lose precision and generate unusual errors that will be captured by the control charts, providing clear evidence of drift. An interesting example of the use of control charts in this regard can be found in [10] where monitoring the residual values of the forecast provided clear evidence of a significant change in the environment due to the emergence of a product with a new technology on the market, which led to the loss of model accuracy.

Given this need for models that adapt to changing environments, the scarcity of forecasting models based on reinforcement learning found in our previous study [1] is surprising, since this is the only machine learning paradigm where learning is done through the continuous interaction and adaptation of the agent with its environment. This is even more surprising considering that, according to [11], the popular models of supervised learning, although important forms of learning, are not suitable for learning from adapting to a changing environment since in these cases it is impractical, if not impossible, to obtain samples that are both correct and representative of all possible situations in which the model must act. Therefore, in adaptive situations, the agent is forced to learn from his or her own experience, which is only possible through reinforcement learning models [11].

To help address the situation described, this study seeks to know the state of the art of the use of reinforcement learning in the field of forecasting the demand for products or services, placing special emphasis on detailing the main problems it addresses, the algorithms used, and the results obtained. Considering the scarcity of articles on this topic found in previous research, this study does not restrict its search to any specific business sector. This study arises from the need to complement and deepen the knowledge about the advances in the field of the use of artificial intelligence in demand forecasting initiated by the authors in a previous study [1]. In this research, no forecasting models were found that use the important paradigm of reinforcement learning, so the present study was necessary to explore this important sector of knowledge. The main contribution of this research is to have demonstrated the enormous potential of reinforcement learning to improve the accuracy of forecasting models in changing and complex environments thanks to its ability to learn and adapt by interacting with the environment. In this sense, we aspire to have contributed to an academic debate that promotes the development of new and innovative models based on the combination of this technology with other state-of-the-art technologies, based on a rigorous examination of the existing scientific information on the subject.

It is important to note that the findings of this study have significant implications for both academia and decision-makers in operations management. First, they suggest considerable limitations to be considered regarding the use of models based on supervised learning in dynamic environments, which can be overcome by combining them with reinforcement learning models. In addition, the results point to the importance of providing forecasting models with mechanisms for detecting significant changes that warn of the need for retraining or updating with emerging data. Finally, the study provides evidence on the effectiveness of reinforcement learning in demand forecasting, which can be used by companies to improve their operational performance.

The rest of the article is structured as follows: in section 2, the fundamental concepts of reinforcement learning are introduced, and the most representative algorithms used in this field are described. Section 3 details the research methodology used, based on the PRISMA approach [12], selected for its suitability to the objectives of the study. Section 4 presents the results obtained from the analysis of the articles and proposes a model for readers' consideration. Section 5 discusses the proposals found in the

analyzed articles and provides the conclusions of the research. Finally, the references used in the study are included.

2. BASIC CONCEPTS

2.1. Reinforcement learning

In reinforcement learning, an agent must act in a territory or environment, usually unexplored, and has to learn what the most beneficial actions are according to his or her own experiences. The environment changes state as the actor acts in it, the set of all possible states that the environment can acquire is called a state space and is usually represented by S . The set of all possible actions that the actor can perform is called the action space and is represented by A . The actor receives a reward or punishment, R , in the form of a signal from the environment each time it changes state depending on the action he has taken. The actor's goal is to learn the actions, according to the state of the environment, that lead to the maximization of rewards over time [11]. The rules by which the actor decides what action to take are called policy. For example, an actor may have an ambitious policy if most of the time he takes an unexplored action in search of a better reward. Conversely, an actor has a conservative policy if most of the time he takes an action that he already knows and that has led to good results in the past [11].

2.2. Main reinforcement learning algorithms

An algorithm can try to build a model that simulates the behavior of the environment and then determine the most appropriate actions to take. If so, this algorithm is called model-based reinforcement learning algorithm. An important such algorithm is Monte Carlo tree search (MCTS), which builds a tree of possible future actions and states through simulations to select the most promising populations [11]. However, an agent can try to learn from its experience without trying to build a model of the environment; if so, the algorithm is a model-free algorithm. In this case the agent might try to calculate how valuable each possible state it can reach is (V), or how valuable each possible action it can perform is (Q). From these values, the agent can select the most convenient action. Algorithms that proceed in this way are called value-based models. The most widespread value-based algorithm is Q-learning (QL) [11]. Finally, a model can try to directly optimize its policy without trying to build models of the environment, or calculate state or action values, in which case the algorithm is called a policy-based model. One of the most important policy-based models is the deterministic policy gradient (PGD) [11], see Table 1.

Table 1. Definitions of the main algorithms

Main algorithms	
QL	This algorithm is based on the construction of a table of Q values for each possible action within each possible state. This table is updated with the results obtained after each action performed by the actor. Among the main advantages of this method are its conceptual simplicity, its low computational cost, and its rapid convergence, i.e., the speed with which it finds the optimal values in the Q table. The main limitation of this model is that it can only handle limited state spaces and discrete action spaces. This is due to the need to construct the Q table, which is not possible in spaces of continuous states and actions [13].
Deep q network (DQN)	This model, instead of a Q table, uses a neural network, usually based on a convolutional neural network (CNN) whose inputs correspond to the values of the state of the environment and the outputs are the Q values for each possible action. This allows the model to simply and efficiently handle unlimited state spaces. The limitations of this model are that due to the output structure of the neural network it can only handle discrete action spaces [13], and in dynamic and complex environments it can fall into an overestimation of Q values [13], [14].
Double deep q network (DDQN)	This model is an improvement of the DQN model to reduce its overestimation problems. To do this, it uses a second neural network that is a copy of the first but does not update its parameters after each action but after every certain period. This allows the model to effectively deal with correlation problems that lead to overestimation of Q values. The limitations of this model are, like the previous model, that it cannot handle continuous action spaces, in addition to retaining certain difficulties with dynamic and complex environments [13], [14].
Deep deterministic policy gradient (DDPG)	This model uses two neural networks, the first called the actor network, which has the values of the state of the environment as inputs and the corresponding action as output. The second network, called the critical network, has the mission of evaluating the quality of the action carried out by the actor network. The inputs of the critical network are the values corresponding to the state plus the action performed by the actor, while their output is the Q value of that action. The advantages of this model are that it can handle unlimited and complex state environments, and that it allows for spaces of continuous action. Its weaknesses are its difficulty of convergence [4], [15] and, like the previous model, overestimations by temporal correlations [15].

3. METHOD

This systematic review of the literature was carried out under the PRISMA methodology, which was created to guarantee the rigor of this type of studies while avoiding possible biases [12]. Additionally, the

selected documents were classified using automatic clustering algorithms to provide an objective classification of the different uses and methods of reinforcement learning in the field of demand forecasting.

3.1. Research questions

As part of the research process, six research questions have been proposed to guide the entire research and to extract and synthesize the knowledge contained in the documents examined. These questions are shown in Table 2.

Table 2. Research questions

Code	Question
Main	How has reinforcement learning been used in the development of demand forecasting models in recent years?
P	What demand forecasting issues or challenges have been addressed with reinforcement learning?
I	What reinforcement learning algorithms have been used for this purpose?
C	What metrics have been used to measure the performance of the proposed models?
O	What is the performance of the reinforcement learning models in relation to the established models?
C	In which business sectors have forecast models based on reinforcement learning been used most frequently?

3.2. Search strategy

For the construction of the search chain, the first two factors of the PICOC methodology were considered: population and intervention. The other factors were not considered to expand the search range. Table 3 shows the search terms related to each of these factors.

Table 3. Search terms

Factor	Description	Search terms	Synonymy
Problem	Demand forecasts for business products and services	"Demand forecasting"	"demand prediction" "demand prognostic" "demand prognosis" "product forecasting"
Intervention	Forecasting using reinforcement learning	"Reinforcement learning"	"reward based learning" "value based learning" "policy based learning" "model based learning" "Q learning" "Deep q network" "DQN" "policy gradients" "Actor Critical Method" "Proximal Policy Optimization" "Trust Region Policy Optimization"

The search terms were combined with Boolean operators to construct the following search string with which the research was conducted in the Scopus, Web of Science (WoS), and IEEE databases:

("demand forecasting" OR "demand prediction" OR "demand prognostic" OR "demand prognosis" OR "product forecasting") AND ("reinforcement learning" OR "reward based learning" OR "value based learning" OR "policy based learning" OR "model based learning" OR "q learning" OR "deep q network" OR "DQN" OR "policy gradients" OR "actor critic method" OR "proximal policy optimization" OR "trust region policy optimization")

For the Google Scholar search, a different search string was used, iteratively constructed according to the results obtained until a set of articles relevant to the research was found. Below is the string used in this search engine:

In title: "Demand Forecasting" + "Reinforcement Learning" + ("Proposed Model" or "Proposed Framework") - "Literature Review"

3.3. Eligibility criteria

To ensure the selection of relevant and high-quality studies for this systematic review, specific inclusion and exclusion criteria were defined. Table 4 details the inclusion criteria. Conversely, Table 5 outlines the exclusion criteria, specifying conditions under which studies are deemed unsuitable for analysis. These tables provide a clear framework to guide the selection process, ensuring the review's rigor and focus.

Table 4. Inclusion criteria

Code	Description
I1	Articles that propose a novel quantitative method for demand forecasting
I2	Articles that make use of "reinforcement learning" within the demand forecast construction process
I3	Empirical articles with models validated with real data from companies
I4	Articles with access to the full text

Table 5. Exclusion criteria

Code	Description
E1	Articles published in languages other than Spanish or English
E2	Articles published before 2018
E3	Documents other than scientific peer-reviewed articles and conference papers
E4	Articles that do not refer to business products or services demand

3.4. Information Sources

The scientific databases Scopus, WoS, and IEEE were chosen as sources of information because they are recognized for their reliability among the academic community. In addition, the Google Scholar search engine was also consulted, however, to guarantee the quality of the articles coming from this source, the Journal Impact Factor (JIF) of the respective publication was checked, see Figure 1.

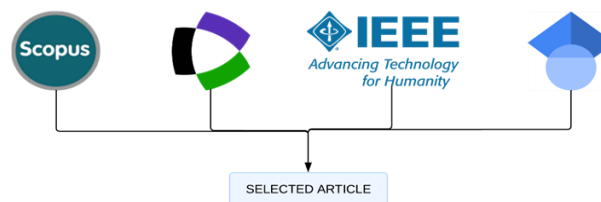


Figure 1. Source of information used in the research

3.5. Article selection process

The research process was carried out in four phases. In the identification phase, search strings were used to locate all articles in the databases that met the specified conditions. In the pre-selection phase, exclusion criteria based on the title and abstract of the articles were applied. During the selection phase, inclusion criteria were also used at the level of title and abstract. Finally, in the inclusion phase, the introduction, methodology and conclusions sections of the articles were reviewed and, applying the inclusion criteria, it was decided whether they would be integrated into the qualitative synthesis. The application of the search strings in the selected databases yielded a total of 438 articles identified: 121 in Scopus, 43 in WoS, 128 in IEEE and 146 in Google Scholar, as can be seen in Figure 2. After removing duplicates, the articles were reduced to 341, after applying the exclusion criteria, 224 articles were eliminated, and 117 remained for the evaluation of the inclusion criteria. After this last evaluation, 93 articles that did not meet at least one criterion were eliminated, leaving a total of 24 articles for inclusion in the qualitative synthesis. Two articles from Google Scholar [16], [17] arrived at this last stage, so the respective JIF was verified, finding that both publications, Mathematics and IEEE Transactions on Smart Grid, belong to the first quartile (Q1) of their respective categories, so both articles were included in the aforementioned qualitative synthesis.

3.6. Automatic grouping

To support the analysis, we sought to classify the selected documents into groups according to their similarity. To carry out this task objectively and free of any bias that could be introduced by the authors, we chose to use a hierarchical agglomerative clustering algorithm. This method was proposed for use in systematic reviews of the literature in [18] and has been used by [1] in a study like the present one, obtaining good results. During the application of this methodology, it was necessary to construct a table of characteristics of the articles, which served as a basis for the calculation of the euclidean distances between them and for the elaboration of the corresponding dendrogram that allowed to visualize the proximity between documents. Finally, the silhouette method was used to determine the optimal number of clusters. The following python libraries were used for this purpose: dendrogram and linkage from scipy.cluster.hierarchy, AgglomerativeClustering from sklearn.cluster, and silhouette_score from sklearn.metrics. The method of measuring distances between clusters was Ward's, as it provided greater distances between groups in the dendrogram than those obtained with the other alternatives.

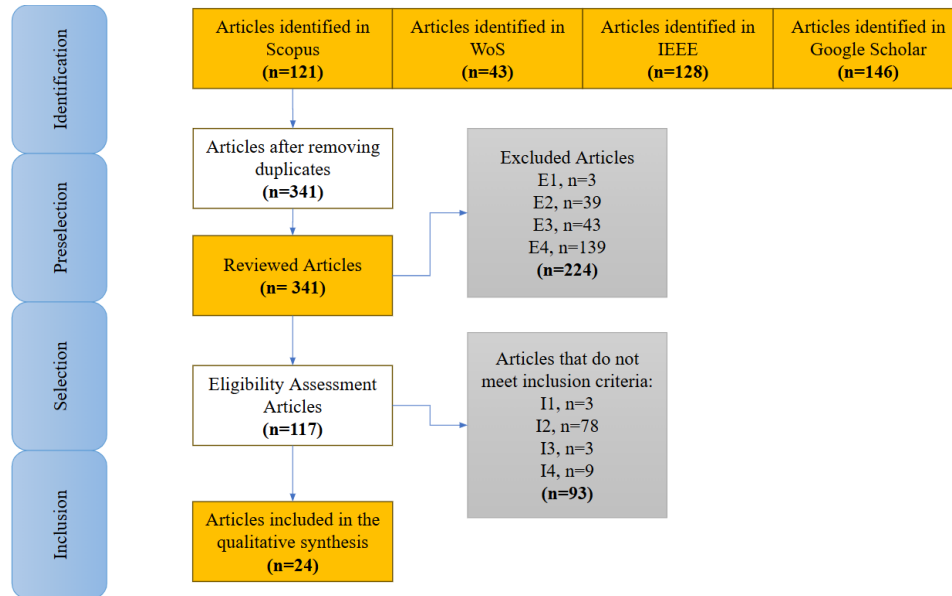


Figure 2. Results of article selection

3.6.1. Hierarchical agglomerative clustering of articles

Five characteristics were determined to be relevant for the classification of the documents: i) whether the reinforcement learning method used is tabular or approximate, which was labeled as "Tabular", ii) whether the reinforcement learning method is based on optimizing the value function of the action or whether, failing that, it tries to directly optimize the policy, which was labeled as "QFunction", iii) if the model uses independent variables in addition to the time series, labeled as "Ivariables", iv) if the main reinforcement learning algorithm used in the proposed model is QL, DQN, DDQN or DDPG. This feature was labeled as "Algorithm" and finally, v) if the objective of reinforcement learning is directly to predict demand (Predict), select the best base predictor (Selection), or find a weighted weight to integrate the base predictors (Weighing), among others. The latter feature was labeled "RLobjective". Documents [19], [20] were excluded from the automatic classification as they do not provide information about the reinforcement learning algorithm used, these documents were grouped directly into an additional cluster. Table 6 shows the resulting table of document features. After converting the categorical variables to dummy variables, the graph of silhouette coefficients and the dendrogram were obtained, which can be seen in Figures 3 and 4. The former reaches its maximum with four clusters, which have been identified on the dendrogram shown.

Table 6. Documents features

IdDoc	Tabulate	QFuntion	Ivariants	Algorithm	RLobjective
1	1	1	1	QL	Selection
2	1	1	1	QL	Selection
3	1	1	1	QL	Weighing
4	0	1	0	DQN	Selection
5	1	1	1	QL	Selection
7	0	1	1	DQN	Select_SDay
8	0	0	0	DDPG	Predict
9	0	1	1	DQN	Opt_Hyper
10	1	1	0	QL	Selection
12	0	0	1	DDPG	Weighing
13	0	0	1	DDPG	Predict
14	0	0	0	DDPG	Weighing
15	0	1	0	DDQN	Selection
16	0	1	0	DQN	Select_SkipL
17	0	0	1	DDPG	Predict
18	0	1	1	DDQN	Selection
19	0	0	1	DDPG	Predict
20	0	0	0	DDPG	Predict
21	0	1	0	DDQN	Classification
22	1	1	0	QL	Selection
23	1	1	0	QL	Weighing
24	1	1	1	QL	Opt_IdleT

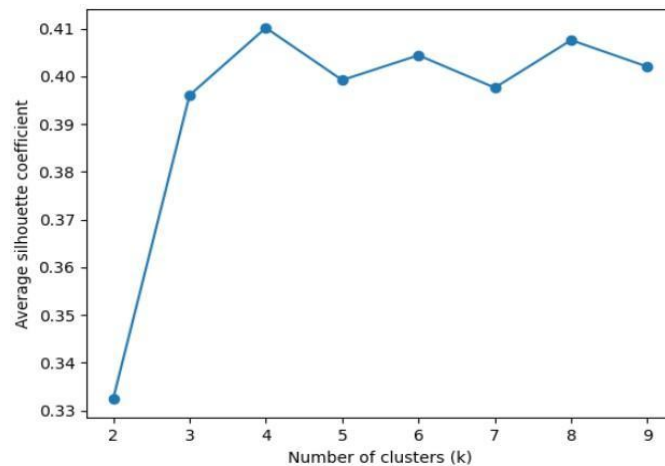


Figure 3. Silhouette method

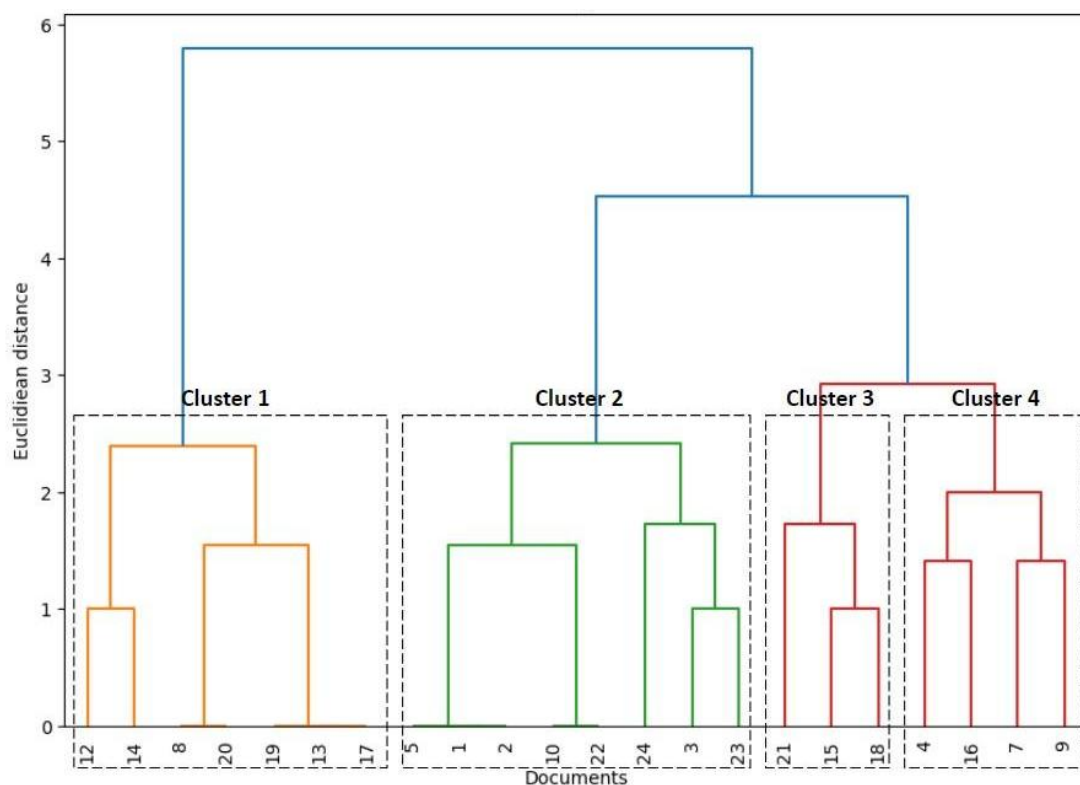


Figure 4. Dendrogram

3.6.2. Clusters

Table 7 (in Appendix) [4]–[7], [15]–[17], [19]–[35] shows the description of each of the clusters obtained with the automatic grouping of articles. The main features that organize these clusters is the algorithm used, which was not obvious before the classification. As evident in the table, the consistency of the clusters is very good, which demonstrates the virtues of automatic grouping.

4. RESULTS AND DISCUSSION

4.1. Results

This section answers the research questions in the light of the analysis of the selected documents while discussing the important aspects related to the results obtained. Main question: how has reinforcement

learning been used in the development of demand forecasting models in recent years? Reinforcement learning models have been used in four main ways in the construction of demand forecasting models:

- To predict demand directly [4]–[6], [15], [21].
- To dynamically select, according to environmental variation, the best predictor from a set of base predictors [7], [16], [17], [24], [25]–[27], [32].
- To integrate, by means of weighted weights, a set of base predictors, according to the variations of the environment [22], [23], [28], [29].
- For hyper-parameter optimization, selection of days and similar segments among the data, and other secondary tasks [31], [33]–[35].

4.1.1. Q1: what demand forecasting issues or challenges have been addressed with reinforcements learning?

Each document analyzed addresses unique problems and proposes different solutions, highlighting the complexities involved in accurately forecasting demand. The challenges identified range from dealing with high volatility and dynamism in demand patterns to managing the intricacies of demand variability across different time scales. We offer a comprehensive overview of the primary challenges and issues discussed in these documents, see Table 8.

Table 8. Results of the keywords corresponding to Q1

Keyword	Input
Demand in highly changing situations	The main problem that reinforcement learning aims to address is forecasting under conditions of high volatility and dynamism of demand [1], [4], [5], [27], [29]. The aim is to build a model that is effective in both peak and normal periods, that does not ignore local variations [17], [24], that works in several scenarios [16], [25] and that adapts quickly to new conditions [20].
Variability on different time scales	Another important challenge, which is sought to be solved with reinforcement learning, is the identification of demand patterns on different time scales that are juxtaposed and generate complicated behaviors [4], [6], [35].
Demand in highly complex situations	Finally, the aim is to use reinforcement learning to address prognostic situations where the high dimensionality of the explanatory variables and possible actions [5], [22]–[24] complicates the model to a high degree.

4.1.2. Q2: what reinforcements learning algorithms have been used for this purpose?

In addressing the various challenges and issues detailed in the preceding section, researchers have employed four key algorithms. Each algorithm offers distinct advantages and is suited to specific types of problems. We mention these four algorithms, highlighting their applications in solving the outlined problems, see Table 9.

Table 9. Results of the keywords corresponding to Q2

Keyword	Input
QL	Primarily used for predictor selection in simple or discretized state spaces.
DQN	Used for various tasks in more complex state spaces such as best predictor selection, hyperparameter optimization, and more.
DDQN	Primarily used for dynamic selection of the best predictor in unlimited state spaces.
DDPG	Mainly used for direct forecasting of demand in unlimited state spaces and for dynamic integration of the predictor base using weighted weights.

4.1.3. Q3: what metrics have been used to measure the performance of the proposed models?

The vast majority of proposed models measure the accuracy of their forecasts with one of the three classic metrics: mean absolute error (MAE), mean square error (MSE), and root mean square error (RMSE). Articles [5], [20], [22] also include the coefficient of determination R² among their metrics. Therefore, the dominant performance metrics among this type of models are error metrics.

4.1.4. Q4: what is the performance of the new proposed models in relation to the established models?

All the proposed models report superior performance to the models taken as a reference in their respective studies or to the baseline predictors taken individually. Two of them [17], [27] even report improvements in forecast accuracy of up to 50% compared to the reference models. This provides strong evidence that including reinforcement learning in forecasting models improves them significantly.

4.1.5. Q5: in which business sectors have forecast models based on reinforcement learning been used most frequently?

Most of the studies come from the electricity sector, 16 of the 24 articles analyzed come from this sector, three articles focus on the transport sector, two on the telecommunications sector, another two on the heat energy sector and only one on the semiconductor distribution sector. It is interesting to note, then, that 23 of the 24 items come from the service sector, especially energy, while only one item comes from the trade sector. This could mean that the use of these types of models is in its early stages and that there are significant opportunities to extend the findings of this study to other sectors such as tourism, retail and manufacturing.

4.2. About the bibliometric analysis

4.2.1. Publication analysis by keywords

For the bibliometric analysis, the VosWierver and Bibliometrix software were used, because good results have been obtained with research related to the systematic review of literature. A total of 438 articles were obtained from the database sources, of which the bibliometric analysis was carried out in the SCOPUS sources. In Figure 5 you can see the most cited keywords in the selected articles, the most used being 'forecasting' followed by 'deep learning', 'learning Systems', and 'reinforcement learning', among others. In Figure 6 you can see the publication trend according to the authors' timeline and the most relevant keywords in the research; For example, the keyword 'forecasting' was most used in 2022 followed by 'deep learning', while the other words were not the most cited in the research.

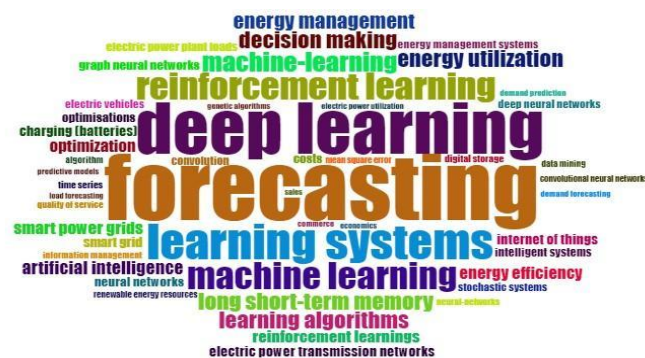


Figure 5. Keywords

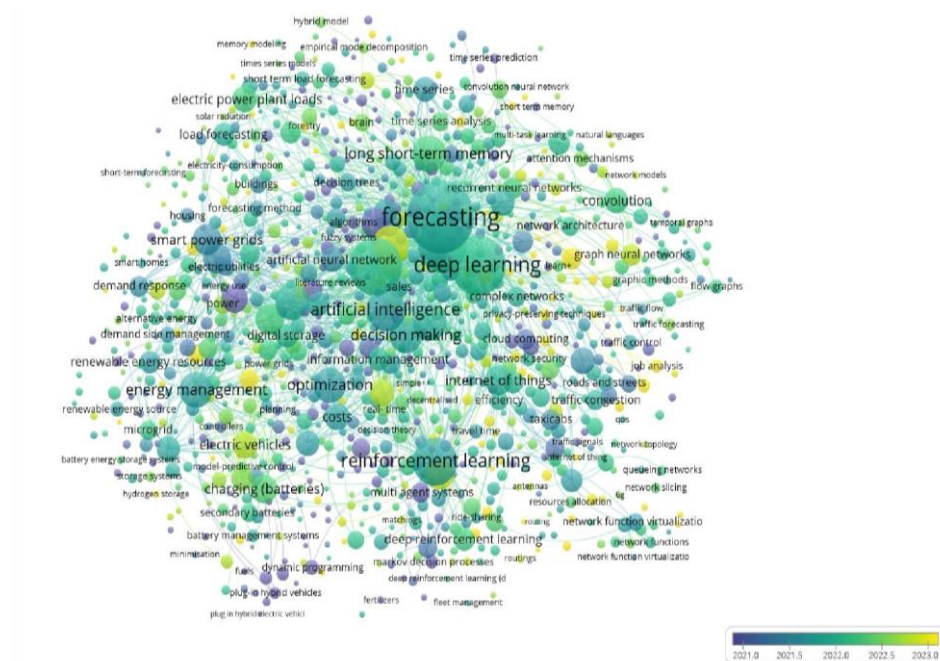


Figure 6. Overlay visualization

4.2.2. Publication trend in different countries

In the Figure 7 it is observed that there is great interest in different countries to investigate the chosen topic, it has a relevant factor in different countries and the citation networks are shown across the edges. First, we see that China is one of the countries, followed by India and the USA, among others.

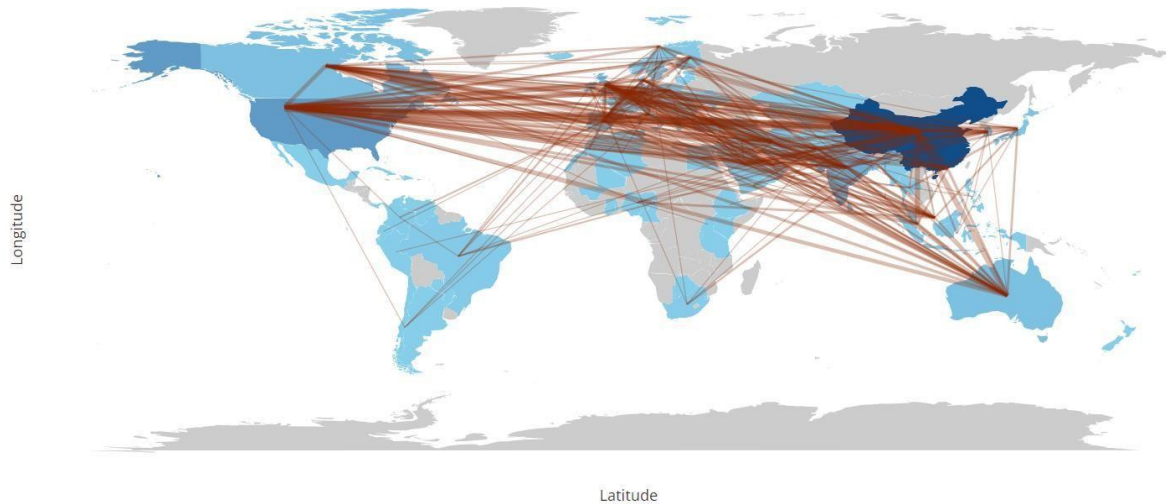


Figure 7. Country collaborations maps

4.3. Proposed model

The results presented above clearly show how reinforcement learning endows forecasting models with a great capacity to adapt to changing environments. However, none of the models found have concept drift detection mechanisms. The latter is especially important for models that use baseline predictors built through supervised learning [8]. To be adaptive, these models need a drift detection mechanism that determines the need to update or retrain the predictors with the new emerging data. This is a complex task, which requires constant monitoring and a high level of understanding and adaptation to what is happening in the environment. In view of the qualities of reinforcement learning illustrated in this study, we propose to improve this type of model through a new use of reinforcement learning: as a mechanism for detecting and controlling drift. We present the proposed model following the scheme designed by [8] for this type of system: memory, learning, loss estimation and change detection.

4.3.1. Memory, learning, and loss estimation

The memory of the proposed model uses the “fixed size sliding windows” method [8] that is, it saves the k most recent data, which means that when a new data arrives it will be saved in the memory and the oldest data will be forgotten. The window size, k , is the same as the size of the training data of the predictor base, built through supervised learning. Regarding the learning mechanism of the model, this consists of a reinforcement learning agent that selects the best predictor from a base of j predictors trained through supervised learning. The algorithm corresponding to this agent has been called reinforcement learning selector (RLS). These predictors are retrained on data stored in memory when determined by the change detection mechanism. Finally, regarding the loss estimate of the model, the average square of error (SME) was chosen for greater sensitivity to increasing errors. The SME will be calculated at each movement of the sliding window and in the presence of concept drift, the predictors are expected to generate larger and larger errors, so the SME will grow rapidly.

4.3.2. Change detection

Change detection is inspired by control charts. As in this [10], the proposed model monitors the sequence of forecast errors generated by each base predictor, but instead of defining upper and lower limits of control, an agent is used to learn from the behavior of the sequence of errors and decide at each step whether to retrain the predictor. The algorithm corresponding to this agent was called reinforcement learning controller (RLC). The state space, action space, and reward function for this agent are detailed:

- Let e_t be the forecast error at time t , for $t=k+1, k+2, k+3, \dots$ (where, k has been defined in the previous lines as the size of the training data of the base predictors).

- Let SME_t be the average square of the errors computed at time t , for $t=k+1, k+2, k+3, \dots$. And let a_t be the action taken by the agent at time t , for $t=k+1, k+2, k+3, \dots$.

The state of the environment at time t is defined as:

$$St = \{et - k + 1, et - k + 2, \dots, et, SME_t - k + 1, SME_t - k + 2, \dots, SME_t, at - k + 1, at - k + 2, \dots, at\}$$

that is, at each time t , the state of the environment consists of the collection of the last k forecast errors, the last k SMEs, and the last k actions taken by the agent.

A is defined as the agent's action space as:

$$A = \{0, 1\}$$

where 0 indicates not training the model, and 1 indicates training the model with the data stored in memory at that moment, i.e., at each moment t , the agent decides on the action a_t that can take the value of 0 or 1 depending on whether it decides to retrain the model or not.

R_t is defined as the reward obtained by the agent at time t as:

$$R_t = \frac{1}{1 + e^{(SME_t - SME_{min})}}$$

where SME_t is the mean of the square of the error computed at time t , and SME_{min} is the smallest known SME in the history of the model. This equation was chosen for the reward for the following reasons:

- Large deviation penalty: the function effectively penalizes large deviations from the minimum SME . When the SME_t is much larger than the SME_{min} , the exponential function grows rapidly, causing the reward to decrease significantly, which discourages actions that result in large errors.
- Stability of learning: the smoothness and continuity of the combination of inverse and exponential functions ensures that rewards change in a gradual and controlled manner. This is crucial for learning stability, as it avoids abrupt changes in rewards that could destabilize the agent's learning process.
- Incentive to reduce the SMS_t below the known minimum: the function provides a larger reward if the agent succeeds in reducing the SMS_t below the known minimum, thus incentivizing the search for better solutions.

4.3.3. Expected reinforcement learning controller behavior and recommended algorithms

The RLC is expected to learn to distinguish noise and outliers since including this type of data in model training must necessarily lead to a worsening of the SME . It is also expected that in the real presence of drift, the agent will learn to detect and include the new data in the training of the models, since if this does not do so, this would lead to a worsening of SME . Finally, it is expected that the agent will learn to distinguish whether to include complex data in the training of predictors according to their impact on the SME . Value-based algorithms are recommended to implement the RLC, since the decision to include or not recent data in the training of predictors has a lot of influence on the value of the future SME , as well as, since the state space is unlimited and complex, but the space of action is discrete, it is recommended to use DQN and even DDQN to avoid the risk of overestimation of Q-values. See Figure 8 for a general outline of the proposed model.

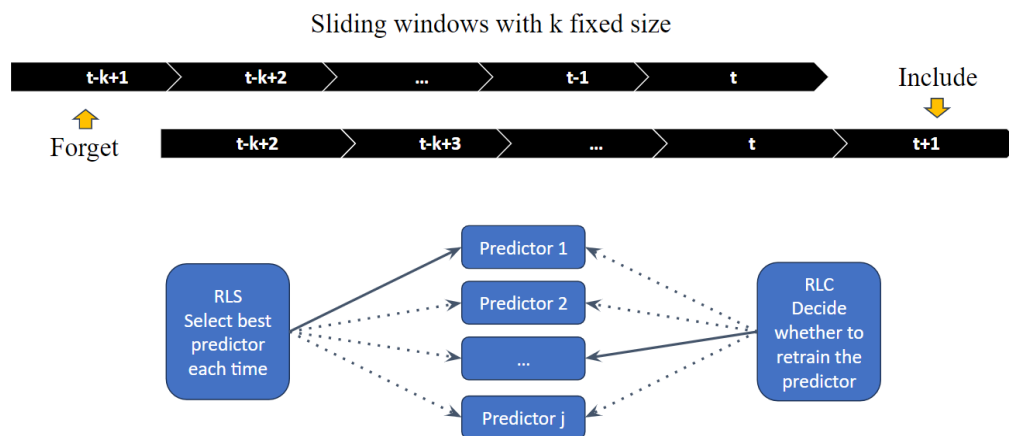


Figure 8. Schematic of the proposed model

4.4. Discussion

One of the most important uses of RL is direct demand forecasting. The greatest similarity among those who propose this use is the unanimous use of the DDPG algorithm [4]–[6], [15], [21], this is because predicting demand directly implies a continuous action space and state spaces of high dimensionality, both situations addressed by the DDPG [13]. Within this group, the authors of [6], [15] differ from the rest in that they only use time series without using other independent variables, even so, the data they handle is complex due to its high variability, which forces them to complement reinforcement learning with devices that facilitate the convergence of the model such as "adaptive early forecasting" in [15] or the decomposition of the time series through "dynamic time warping" proposed in [6] to reduce its complexity. The rest of the authors use independent variables related to climate [4], [21] and spatiotemporal data [5], which adds complexity to the models. In this subgroup [4] coincides with [6] in the decomposition of the time series, this time using "wavelet transformation", to reduce the difficulties of DDPGs in learning in the face of large fluctuations. A limitation of all these studies is, paradoxically, their exclusive focus on the DDPG algorithm: according to [13] there are seventeen other reinforcement learning algorithms capable of handling unlimited state spaces and continuous action spaces. It is therefore recommended that future research explore the benefits of other algorithms for the task of direct forecasting in complex and dynamic environments.

On the other hand, the most frequent use of reinforcement learning is for the dynamic selection of the best predictor from a pool of predictors. All proponents make use of some sort of value-based algorithm for this task. The majority, [16], [17], [25]–[27] propose the use of QL for this purpose, while another group of authors propose neural networks [7], [24], [32]. Among those who propose the use of QL for the selection of the best predictor, we have [16], [25] who use only time series, while the authors [17], [26], [27] include in their model independent variables such as econometric variables [26] and climate-related variables [1], [17]. Ingeniously, to deal with the limitations of QL, these authors handle these variables at the level of the predictor base while keeping the state space for selection very simple: [26] includes the last three chosen predictors and the last three discretized demands in this space, while [17], [27] use predictor rankings in their state spaces. A diametrically opposed approach is followed by proponents of DQN [7] and DDQN [24], [32] for the selection of the best predictor, they allow themselves complex state spaces thanks to the use of neural networks. The main limitation of all these selection models is that they do not have any drift detection mechanism that indicates the need to update or retrain the baseline predictors. The model we propose in this article approaches this task with reinforcement learning. It is therefore recommended that future studies explore this avenue by verifying with empirical data the efficiency and effectiveness of what is proposed here and its possible variants.

Regarding the use of reinforcement learning to assemble a set of predictors using weighted weights, we find two groups of authors: On the one hand, there are ones propose using DDPG [22], [23] given that state spaces are extensive and complex; and on the other hand, there are those who propose using QL [28], [29] who, due to the limitations of the algorithm, are forced to discretize both the state space (the explanatory variables) and the action space (the weighted weights), this goes against the recommendation of [13] who indicates that the algorithm that adapts to the needs of the environment should be chosen and not try to simplify the environment to adapt it to the algorithm. Even so, these studies report superior performance than the models taken as a reference. However, all these models share the same limitation as selection models: they don't have a mechanism for determining when to retrain their baseline predictors.

Regarding the problems that are addressed with reinforcement learning, we have the authors who address the volatility and dynamism of demand with direct prediction using DDPG [4], [5], but we also have those who face it with the simplest QL both by integrating predictors [29] and selecting the best of them [27]. The search for a model that can be used for peak periods, normal periods, local variations and in several scenarios has been approached by the authors mainly through the selection of predictors through QL [16], [17], [25], although we also found some that have done so with more complex models such as DDQN [24]. As for the problem of demand variability at different time scales, this has been addressed by the authors exclusively by direct prediction with DDPG combined with time series decomposition techniques [4], [6]. Finally, demand prediction in highly complex situations has been addressed with DDPG both with direct prediction [5] and with weighted integration of predictors [22], [23] as well as with DDQN for the selection of predictors [24]. Although the great capacity of these algorithms to adapt to changing situations has been demonstrated, again the limitation of these models, except for those of direct prediction with DDPG, is that they lack a way to evaluate whether the intensity of the fluctuations has caused concept drift, and therefore, if it is necessary to update the predictors of supervised learning.

Another important result is the fact that the algorithms used for demand forecasting are concentrated on the four identified: QL, DQN, DDQN, and DDPG. This is surprising since there are a wide variety of unused algorithms [13]. This could be an indicator that the application of reinforcement learning to the field of demand forecasting is in its early stages, which is why research is focused on the most well-known

algorithms. In any case, it is highly recommended that future studies focus on investigating the effectiveness and efficiency of other reinforcement learning algorithms to expand the field of application.

No less surprising is the high concentration of studies in the energy sector, with a few studies in the transport and telecommunications sector, and only one in the trade sector. This would also indicate that this field of study would be in its early stages, being cultivated mainly by sectors with high demand volatility such as the electric power sector. If this is the case, there would be the possibility of extending these studies to other areas such as tourism, industry and retail, which are also characterized by their high volatility. Another interpretation for the underdevelopment of the field could be that there are barriers to a more widespread use of reinforcement learning, such as its high computational cost, convergence problems, and long times to reach optimal performance. If this is the case, studies on the application of QL, with its conceptual simplicity and rapid convergence, could be the spearhead for the expansion of reinforcement learning to other sectors. As for more complex models, studies on complements such as time series decomposition [4], [6] or algorithms to accelerate convergence such as adaptive early forecasting in [15], could be vital for the expansion of the field.

5. CONCLUSION

The present research was able to answer all the research questions posed. Regarding the question of how reinforcement learning has been used in demand forecasting, it was established that it has been used mainly for the dynamic selection of the best predictor from a base of predictors. It has also been used to directly predict demand and integrate a base of weight-weighted predictors. Regarding the problems and challenges that have been addressed with reinforcement learning, it was determined that it has been used for forecasts in highly changing situations, to adapt to variability on different time scales and for forecasts in situations of high complexity of the environment. Regarding the question of which are the most used algorithms, it was specified that these are QL, DQN, DDQN, and DDPG. Regarding the question of what metrics have been used to measure the performance of the models, it was pointed out that the most used are the classic MAE, mean absolute percentage error (MAPE), RMSE and R^2 . Regarding the question about the performance of the models that use reinforcement learning, it was established that they achieved a higher performance than the models with which they were compared, some even reported a 50% improvement in the error metrics in relation to them. Finally, when asked about the business sectors where this type of forecast has been developed, it was determined that it has mainly been the electric power sector, with other sectors having very little representation, such as the commerce sector, which only has one study. Additionally, this study has proven that the proposed models, especially those that use predictors developed with supervised learning, do not have concept drift detection mechanisms, so they cannot know when it is necessary to retrain their base models. To address this problem, a new use for reinforcement learning for drift detection is conceptually proposed. Finally, this study has revealed the need for future research that explores the large number of other algorithms of RL available, in addition to the widely used QL, DQN, DDQN, and DDPG. Likewise, the need for further research on concept drift detection mechanisms, such as the one proposed here, and on the use of RL for demand forecasts in other sectors such as tourism, industry and retail trade, became evident.

APPENDIX

Table 7. Clusters descriptions

Cluster	Definition
Cluster 1: the policy gradient group.	<p>This is a group of seven papers that use the DDPG algorithm, of which five, [4]-[6], [15], [21] from the electric power, heat, telecommunications and transport sectors, use it to directly predict demand; while the other two [22], [23], from the electric power and heat sectors, respectively, use it to dynamically determine the most appropriate weighted weights to integrate a database of predictors.</p> <p>The main problems that these papers try to solve with the use of the DDPG are the variability at different time scales [4], [6], the nonlinearity, volatility, and dynamism of demand fluctuations in their respective sectors [4], [5] and the limitations of tabular methods of reinforcement learning, such as QL to handle spaces of state and continuous actions of great dimensionality [5], [13], [23].</p> <p>The challenges for demand forecasting in the situations addressed by these articles are, firstly, the high granularity of the data, which is hourly in all but one of them [6] where it is even greater: at the level of seconds; secondly, the inclusion in four of them of independent variables related to climate such as temperature or radiation. This situation generates very high-dimensional state spaces. On the other hand, reinforcement learning algorithms are used directly to predict demand or assign weighted weights to the predictor base, in both cases the space for action is therefore continuous. These two conditions, unlimited state spaces and continuous action spaces, make the DDPG algorithm the best choice [13].</p> <p>Despite the above, the use of the DDPG algorithm requires considering its inherent weaknesses, such as the instability of convergence or the high temporal correlation that leads to overestimating the value of the policy over</p>

Table 7. Clusters descriptions (*continued*)

Cluster	Definition
	time, in this sense articles [4], [15] are interesting. In the first, it introduces two critique networks that work asynchronously, i.e. they are updated in different periods, to effectively deal with the high temporal correlation, and introduces an adaptive method of early forecasting that prevents the agent from exploring actions that are not at all beneficial, shortening the training time of the model and preventing the reduction of its accuracy. The second article uses wavelet decomposition to reduce time series to less complex sub-time series on which DDPG is applied. This alleviates the difficulty of convergence that the DDPG has on highly fluctuating data. Regarding the results of the application of these models, all studies reported an improvement over the forecasts of the baseline predictors taken individually and over other state-of-the-art models taken as a reference. The most used metrics were RMSE, ASM, and MAE.
Cluster 2: the q learning group.	It is made up of eight papers whose main feature is the use of the QL algorithm mainly for dynamic selection of the best predictor, which can be seen in five papers [16], [17], [25]-[27]; although it is also used for the integration of predictors using weighted weights, which can be found in two documents [28], [29]. These articles all come from the electricity sector unlike the last member of this cluster, the document [30], which comes from the transport sector that uses QL during the forecasting process to simulate the behavior of the carriers and understand the dynamics of the system. The problem that the articles in this cluster try to solve is to find a technique that works in several scenarios [16], [25] and under changing situations [27], [29], and that solves problems typical of traditional forecasting methods such as improving the overall accuracy of the forecast by ignoring local variations [17], recursive errors when using independent variable forecasts to make forecasts of dependent variables [26] or errors entered by data outliers [28]. To use QL, these models are required to define their state and action spaces discretely. This can be clearly seen in [30] where the state space is "n" stations and their need for dispatch or reception of bicycles. Likewise, the space of action of the agent (the carrier) consists of choosing whether to remain inactive or to perform a transport service. Another example is provided by studies [16], [25] where the task of QL consists of choosing between two predictors, ANN and RNN, according to the performance of each with respect to the observed data. In the latter cases, the action space consists of only two possible actions: choosing one algorithm or another, and the state space consists of a brief set of the most recent selections and their results. In the two cases where QL selects weighted weights to integrate the predictors [28], [29], the authors are forced to normalize the data, that is, convert them to values between zero and one, and then convert them into discrete quantities to represent the state spaces. Similarly, the possible weighted weights for each predictor are discrete quantities previously set to represent the share space in a discrete way. The latter goes against the recommendation of [13] that it is preferable to choose an appropriate algorithm for the environment rather than trying to simplify it to fit the algorithm, but despite this, all the documents in these sections report superior performance of the proposed models against individual algorithms or other reference algorithms.
Cluster 3: the double q network group.	This group is made up of three papers that use the DDQN algorithm, one [31], from the telecommunications sector, to identify similar consumer segments for forecasting purposes; and two from the electricity sector [24], [32] for dynamic selection of the best predictor. The problems that these documents attempt to solve are varied: [31] it tries to accurately classify consumers in order to make adequate forecasts, [32] it tries to overcome the scarcity of data by making use of analogous data obtained in situations similar to the one it is trying to forecast, and [24] it seeks to find a method that is effective both in demand peaks and in more normal periods. The action space of these models is discrete: selection of the best predictor or the best segment, but the state space is unlimited by the multiple independent variables they include. If, in addition to this configuration, there are reasons to suspect overestimation of Q values, then the DDQN algorithm is suitable [13]. This occurs in complex environments where short, very wide fluctuations (peaks) alternate with regular fluctuations such as the one described in [24].
Cluster 4: the q network group.	This cluster is made up of four documents that use the DQN algorithm for different purposes within the forecasting process: [33] uses it to optimize the hyperparameters of the predictor model that is based on an LSTM network, [34] to select from the past data a day similar to the day that is the subject of the forecast, [35] to find demand patterns in different time horizons, and finally [7] to dynamically select the best predictor. This last article comes from the semiconductor distribution sector as opposed to the first three which come from the electricity sector. The problems that these articles attempt to solve are also varied: [33] attempts to overcome the lack of a rigorous method to optimally adjust the hyperparameters that have a high incidence on the accuracy of the forecast. [34] tries to find a method to find days in the past data like the one you are trying to forecast, which is especially complicated due to the nonlinear relationship between the large number of independent variables and the dependent variable. [35] addresses the problem of identifying demand patterns on different time scales, while finally [22] seeks an automatic method to select the best predictor for different products in different markets and influenced differently by the various independent variables. As in the previous cluster, the state space is unlimited, and the action space is discrete, which makes the DQN the simplest candidate to address these problems [13].
Cluster 5: other studies.	In this group, the two articles that used reinforcement learning for their forecasting model were combined, but do not specify what type of algorithm was used. The first of these [19], from the transport sector, uses reinforcement learning to weight and integrate base predictors; while the second [20] uses reinforcement learning to predict electricity consumption and trends from various combinations of real and simulated measurements. The problem that the first paper [20] tries to solve is to find a forecasting model that is capable of quickly adapting to new conditions, while the second [19] tries to overcome the scarcity and variability of monitored data to forecast electricity consumption at the city level.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the Universidad Tecnológica del Perú and the Graduate School of the Universidad Nacional Mayor de San Marcos for providing the academic context in which this

study was conducted. It is important to note, however, that no external funding or research contract was granted for the completion of this research.




REFERENCES

- [1] J. R. N. Villar and M. A. C. Lengua, "A Systematic Review of the Literature on the Use of Artificial Intelligence in Forecasting the Demand for Products and Services in Various Sectors," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 3, pp. 144–156, 2024, doi: 10.14569/IJACSA.2024.0150315.
- [2] B. Rolf, I. Jackson, M. Müller, S. Lang, T. Reggelin, and D. Ivanov, "A review on reinforcement learning algorithms and applications in supply chain management," *International Journal of Production Research*, vol. 61, no. 20, pp. 7151–7179, Oct. 2023, doi: 10.1080/00207543.2022.2140221.
- [3] M. Muth, M. Lingenfelder, and G. Nufer, "The application of machine learning for demand prediction under macroeconomic volatility: a systematic literature review," *Management Review Quarterly*, pp. 1–44, Jun. 2024, doi: 10.1007/s11301-024-00447-8.
- [4] R. Zhang *et al.*, "Short-Term Power Load Forecasting Based on Wavelet Transform and Deep Deterministic Policy Gradient," *IET Conference Proceedings*, no. 27, pp. 158–163, Feb. 2022, doi: 10.1049/icp.2023.0092.
- [5] S. Qiao *et al.*, "An three-in-one on-demand ride-hailing prediction model based on multi-agent reinforcement learning," *Applied Soft Computing*, vol. 149, p. 110965, Dec. 2023, doi: 10.1016/j.asoc.2023.110965.
- [6] J. Ouyang, K. Zhang, H. Zheng, F. Wu, and X. Huang, "Cross-Domain Complementarity and Multi-Time Scale Fusion Based Resource Demand Prediction for Mobile Vehicles," in *2023 IEEE 23rd International Conference on Communication Technology (ICCT)*, IEEE, Oct. 2023, pp. 1019–1024, doi: 10.1109/ICCT59356.2023.10419701.
- [7] C. F. Chien, Y. S. Lin, and S. K. Lin, "Deep reinforcement learning for selecting demand forecast models to empower Industry 3.5 and an empirical study for a semiconductor component distributor," *International Journal of Production Research*, vol. 58, no. 9, pp. 2784–2804, May 2020, doi: 10.1080/00207543.2020.1733125.
- [8] J. Gama, I. Zliobaite, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Surveys*, vol. 46, no. 4, pp. 1–37, Apr. 2014, doi: 10.1145/2523813.
- [9] V. Cerqueira, L. Torgo, F. Pinto, and C. Soares, "Arbitrage of forecasting experts," *Machine Learning*, vol. 108, no. 6, pp. 913–944, Jun. 2019, doi: 10.1007/s10994-018-05774-y.
- [10] A. Durmusoglu, "Updating technology forecasting models using statistical control charts," *Kybernetes*, vol. 47, no. 4, pp. 672–688, Mar. 2018, doi: 10.1108/K-04-2017-0144.
- [11] R. Sutton and A. Barto, "Reinforcement Learning: An Introduction", second edition, in *Adaptive computation and machine learning series*. Cambridge, Massachusetts: The MIT Press, 2018.
- [12] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement," *Journal of clinical epidemiology*, vol. 62, no. 10, pp. 1006–1012, Oct. 2009, doi: 10.1016/j.jclinepi.2009.06.005.
- [13] F. Almahamid and K. Grolinger, "Reinforcement Learning Algorithms: An Overview and Classification," in *Canadian Conference on Electrical and Computer Engineering*, IEEE, Sep. 2021, pp. 1–7, doi: 10.1109/CCECE53047.2021.9569056.
- [14] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-Learning," in *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, 2016, pp. 2094–2100, doi: 10.1609/aaai.v30i1.10295.
- [15] W. Zhang, Q. Chen, J. Yan, S. Zhang, and J. Xu, "A novel asynchronous deep reinforcement learning model with adaptive early forecasting method and reward incentive mechanism for short-term load forecasting," *Energy*, vol. 236, pp. 1–14, Dec. 2021, doi: 10.1016/j.energy.2021.121492.
- [16] M. Zulfiqar, N. F. Alshammari, and M. B. Rasheed, "Reinforcement Learning-Enabled Electric Vehicle Load Forecasting for Grid Energy Management," *Mathematics*, vol. 11, no. 7, pp. 1–20, 2023, doi: 10.3390/math11071680.
- [17] C. Feng, M. Sun, and J. Zhang, "Reinforced Deterministic and Probabilistic Load Forecasting via Q -Learning Dynamic Model Selection," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1377–1386, Mar. 2020, doi: 10.1109/TSG.2019.2937338.
- [18] C. Tauchert, M. Bender, N. Mesbah, and P. Buxmann, "Towards an integrative approach for automated literature reviews using machine learning," in *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2020, pp. 762–771, doi: 10.24251/hicss.2020.095.
- [19] L. A. H. Hassan, H. S. Mahmassani, and Y. Chen, "Reinforcement learning framework for freight demand forecasting to support operational planning decisions," *Transportation Research Part E: Logistics and Transportation Review*, vol. 137, pp. 1–20, May 2020, doi: 10.1016/j.tre.2020.101926.
- [20] S. Maki *et al.*, "A deep reinforced learning spatiotemporal energy demand estimation system using deep learning and electricity demand monitoring data," *Applied Energy*, vol. 324, p. 119652, Oct. 2022, doi: 10.1016/j.apenergy.2022.119652.
- [21] C. Wang, L. Zheng, J. Yuan, K. Huang, and Z. Zhou, "Building Heat Demand Prediction Based on Reinforcement Learning for Thermal Comfort Management," *Energies*, vol. 15, no. 21, pp. 1–20, Oct. 2022, doi: 10.3390/en15217856.
- [22] C. Huang *et al.*, "DearFSAC: A DRL-based Robust Design for Power Demand Forecasting in Federated Smart Grid," in *Proceedings - IEEE Global Communications Conference, GLOBECOM*, IEEE, Dec. 2022, pp. 5279–5284, doi: 10.1109/GLOBECOM48099.2022.10001127.
- [23] J. Sun, M. Gong, Y. Zhao, C. Han, L. Jing, and P. Yang, "A hybrid deep reinforcement learning ensemble optimization model for heat load energy-saving prediction," *Journal of Building Engineering*, vol. 58, p. 105031, Oct. 2022, doi: 10.1016/j.job.2022.105031.
- [24] W. Pannakkong, V. T. Vinh, N. N. M. Tuyen, and J. Buddhakulsomsiri, "A Reinforcement Learning Approach for Ensemble Machine Learning Models in Peak Electricity Forecasting," *Energies*, vol. 16, no. 13, pp. 1–20, Jul. 2023, doi: 10.3390/en16135099.
- [25] M. Dabbaghjamesh, A. Moeini, and A. Kavousi-Fard, "Reinforcement Learning-Based Load Forecasting of Electric Vehicle Charging Station Using Q-Learning Technique," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 6, pp. 4229–4237, Jun. 2021, doi: 10.1109/TII.2020.2990397.
- [26] P. Y. Yin and C. H. Chao, "Automatic selection of fittest energy demand predictors based on cyber swarm optimization and reinforcement learning," *Applied Soft Computing Journal*, vol. 71, pp. 152–164, Oct. 2018, doi: 10.1016/j.asoc.2018.06.042.
- [27] C. Feng and J. Zhang, "Reinforcement Learning based Dynamic Model Selection for Short-Term Load Forecasting," in *2019 IEEE Power and Energy Society Innovative Smart Grid Technologies Conference, ISGT 2019*, IEEE, Feb. 2019, pp. 1–5, doi: 10.1109/ISGT.2019.8791671.
- [28] J. Wang, H. Liu, G. Zheng, Y. Li, and S. Yin, "Short-Term Load Forecasting Based on Outlier Correction, Decomposition, and




- Ensemble Reinforcement Learning,” *Energies*, vol. 16, no. 11, pp. 1–16, May 2023, doi: 10.3390/en16114401.
- [29] M. Ma, B. Jin, S. Luo, S. Guo, and H. Huang, “A novel dynamic integration approach for multiple load forecasts based on Q-learning algorithm,” *International Transactions on Electrical Energy Systems*, vol. 30, no. 7, pp. 1–14, Jul. 2020, doi: 10.1002/2050-7038.12146.
- [30] Y. Guo, J. Li, L. Xiao, H. Allaoui, A. Choudhary, and L. Zhang, “Efficient inventory routing for Bike-Sharing Systems: A combinatorial reinforcement learning framework,” *Transportation Research Part E: Logistics and Transportation Review*, vol. 182, p. 103415, Feb. 2024, doi: 10.1016/j.tre.2024.103415.
- [31] X. Huang, W. Wu, and X. S. Shen, “Digital Twin-Assisted Resource Demand Prediction for Multicast Short Video Streaming,” in *Proceedings - International Conference on Distributed Computing Systems*, IEEE, Jul. 2023, pp. 967–968, doi: 10.1109/ICDCSS57875.2023.00095.
- [32] Y. Fu, D. Wu, and B. Boulet, “On the Benefits of Transfer Learning and Reinforcement Learning for Electric Short-term Load Forecasting,” in *2022 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing & Communications (GreenCom) and IEEE Cyber, Physical & Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics)*, 2022, pp. 195–203, doi: 10.1109/iThings-GreenCom-CPSCom-SmartDataCybermatics55523.2022.00020.
- [33] N. T. N. Anh, D. T. Dat, and L. A. Ngoc, “Reinforcement learning for optimization hyperparameters of Long Short-Term Memory applied to Electricity load forecasting,” in *Proceedings of the 2021 IEEE International Conference on Machine Learning and Applied Network Technologies, ICMLANT 2021*, IEEE, Dec. 2021, pp. 1–6, doi: 10.1109/ICMLANT53170.2021.9690555.
- [34] R. J. Park, K. B. Song, and B. S. Kwon, “Short-term load forecasting algorithm using a similar day selection method based on reinforcement learning,” *Energies*, vol. 13, no. 10, p. 2640, May 2020, doi: 10.3390/en13102640.
- [35] X. Guo, Y. Jiang, L. Li, G. Fu, and S. Yao, “Short-term power load forecasting based on DQN-LSTM,” in *Proceedings of the 34th Chinese Control and Decision Conference, CCDC 2022*, IEEE, Aug. 2022, pp. 855–860, doi: 10.1109/CCDC55256.2022.10034391.

BIOGRAPHIES OF AUTHORS



José Rolando Neira Villar    is a professor at the Universidad Tecnológica del Peru (UTP) and at CENTRUM, the postgraduate school of the Pontificia Universidad Católica del Peru. He has a degree in industrial engineering from the National University of Callao, and a Master in Business Administration from the Pontificia Universidad Católica del Perú, he is currently a doctoral student in systems engineering and computer science at the National University of San Marcos. He works on the use and applications of artificial intelligence in operations management and supply chains. He can be contacted at email: jose.neirav@unmsm.edu.pe.



Miguel Angel Cano Lengua    is a Ph.D. Engineering of Systems and Computer Science from the Universidad Nacional Mayor de San Marcos, a Master's in systems engineering from the Universidad Nacional del Callao (UNAC), a profesor at the Universidad Tecnológica del Peru (UTP) and teaching researcher in the Universidad Nacional Mayor de San Marcos (UNMSM), has a degree in Mathematics. He works on continuous optimization, artificial intelligence algorithms, conical programming, numerical methods, methodology, and software design. He can be contacted at email: mcanol@unmsm.edu.pe.